Imputace nulových hodnot v metabolomice

Alžběta Gardlo[a], Matthias Templ[b], Karel Hron[c], Peter Filzmoser[b]

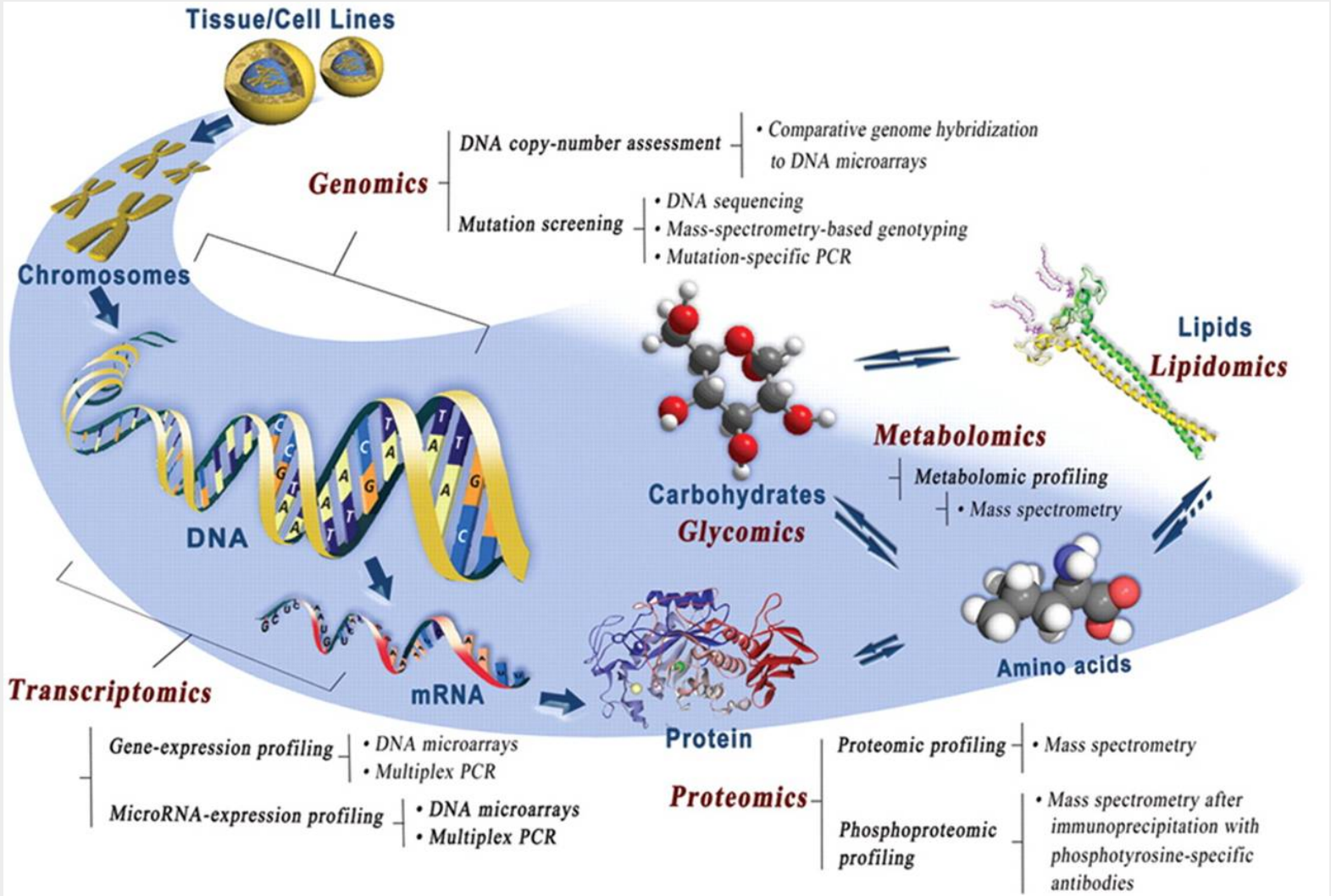
alzbetagardlo@gmail.com

[a] Laboratoř metabolomiky, Ústav molekulární a translační medicíny, Přírodovědecká fakulta, UPOL, Fakultní nemocnice Olomouc;
[b] Vienna University of Technology, Austria;
[c] Přírodovědecká fakulta, UPOL.

Robust, 13.9. 2016

Obsah

Metabolomika



(Wu et al., 2011)

Pivotové isometrické logratio (ilr) souřadnice

- Chceme vytvořit ortonormální bázi vzhledem k Aitchisonově geometrii, ve které první ilr souřadnice vysvětluje veškerou důležitou informaci o zvolené složce.
- Máme kompoziční matici $\mathbf{X}_{n\times D} = (\mathbf{x}_1, \ldots, \mathbf{x}_D)$. Přeuspořádaná kompozice s l-tým prvkem, $l = 1, \ldots, D$, posunutým na první pozici je označena jako $\mathbf{X}^{(l)} = (\mathbf{x}_l, \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_D) = (\mathbf{x}_1^{(l)}, \mathbf{x}_2^{(l)}, \ldots, \mathbf{x}_l^{(l)}, \mathbf{x}_{l+1}^{(l)}, \ldots, \mathbf{x}_D^{(l)})$.

Pivotové ilr souřadnice

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \ldots, D-1. \quad (1)$$

Druhy nulových hodnot

- **Chybějící hodnoty**
  - Hodnota chybí z nějakého důvodu - nelze změřit, respondent neodpověděl na otázku.
  - Nahrazení rozumnou hodnotou.
  - V metabolomice se nevyskytují často.

- **Zaokrouhlené nuly**
  - Vznikají zaokrouhlováním dat blízkých nule nebo díky tzv. **detekčnímu limitu přístroje**.
  - Citlivost každého přístroje má své limity (detekční limit - DL) - hodnoty pod DL jsou vyhodnoceny jako nula, i když by měly být přítomny nějaké koncentrace.
  - Je třeba nahradit s ohledem na DL.
  - Časté v metabolomice, zejména při použití tzv. necíleného přístupu.

Imputace zaokrouhlených nul

- Většina současných statistických metod není schopna pracovat s nulovými hodnotami $\rightarrow$ potřeba imputace.
- Současné metody nahrazování:
  - Nahrazení nulových hodnot 2/3 limitu detekce nebo jinou vhodně zvolenou konstantou - často užívané, ale ignoruje mnohorozměrnou strukturu dat a podhodnocuje kovarianční strukturu.
  - Metoda založená na k nejbližších sousedech - mnohorozměrná, ale pořád ne zcela ideální.

Imputace zaokrouhlených nul - přístupy

- Balíček *zCompositions* v softwaru R.
- **Multiplicative replacement (mult repl)** - nahrazení pomocí části DL (např. 2/3 DL).
- **Multiplicative log-normal replacement (mult lognorm)** - nahrazení nul s využitím multiplikativního lognormálního rozdělení.
- **Multiplicative Kaplan-Meier smoothing spline replacement (mult KMSS)** - nahrazení pomocí geometrického průměru náhodného výběru z kubické vyhlazovací funkce (odpovídá inverzi Kaplan-Meierovy EDF).
- **Log-ratio data augmentation algorithm (lr da)** - využití Markov chain Monte Carlo přístupu pro aditivní logratio (alr) souřadnice.

Imputace zaokrouhlených nul - přístupy

- **Additive log-ratio EM algorithm (lr em)** - postupné využití EM algoritmu pro alr souřadnice.
- **PLS** - využití pivotových ilr souřadnic a metody dílčích nejmenších čtverců - bere v úvahu kompoziční podstatu dat i existenci DL (více v posteru).
- **Pre-selection of variables and model-based replacement of rounded zeros (method varOLS)** - využívá variační matici pro výběr proměnných a redukci dimenze dat.

Validační kritéria

1 Average difference in covariance structure (ADCS)

$$ADCS = \sqrt{\frac{1}{(D-1)^2}\sum_{i=1}^{D-1}\sum_{j=1}^{D-1}\left(s_{ij}-s_{ij}^*\right)^2} = \frac{1}{D-1}\|\mathbf{S}-\mathbf{S}^*\|_F,$$

kde $*$ označuje imputovanou matici, $\mathbf{S}$ je výběrová kovarianční matice, $\|\cdot\|_F$ je Frobeniova maticová norma.

2 Compositional error deviation (CED)

$$\frac{\frac{1}{n_M}\sum_{k\in M} d_a(\mathbf{x}_k, \mathbf{x}_k^*)}{\max\limits_{\{\mathbf{x}_i,\mathbf{x}_j\in\mathbf{X}\}}\{d_a(\mathbf{x}_i,\mathbf{x}_j)\}},$$

Aitchisonova vzdálenost dvou kompozic $\mathbf{x}$ a $\tilde{\mathbf{x}}$:

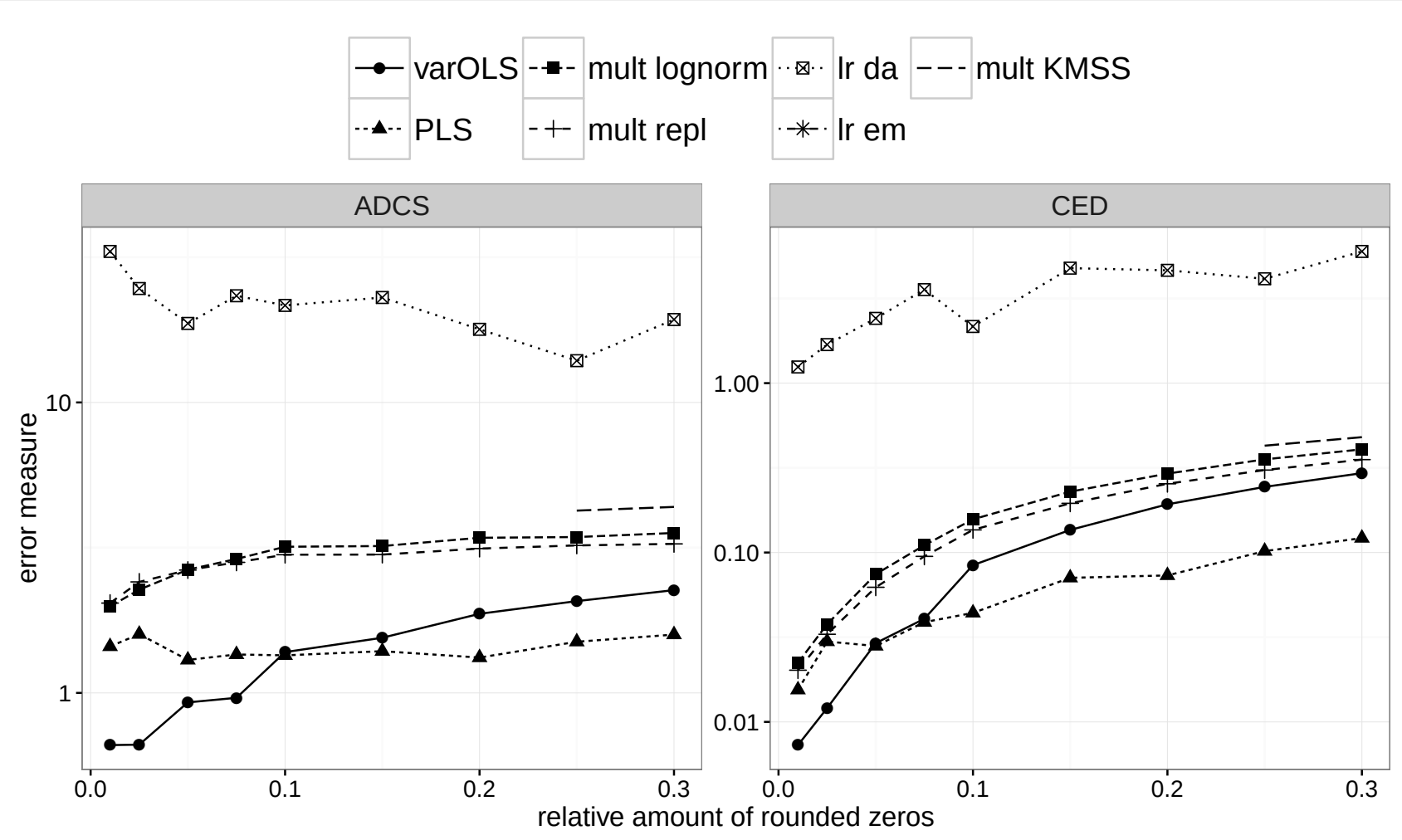
$$d_A(\mathbf{x},\tilde{\mathbf{x}}) = \left[\frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\left(\log\frac{x_i}{x_j}-\log\frac{\tilde{x}_i}{\tilde{x}_j}\right)^2\right]^{1/2}.$$

Simulační studie

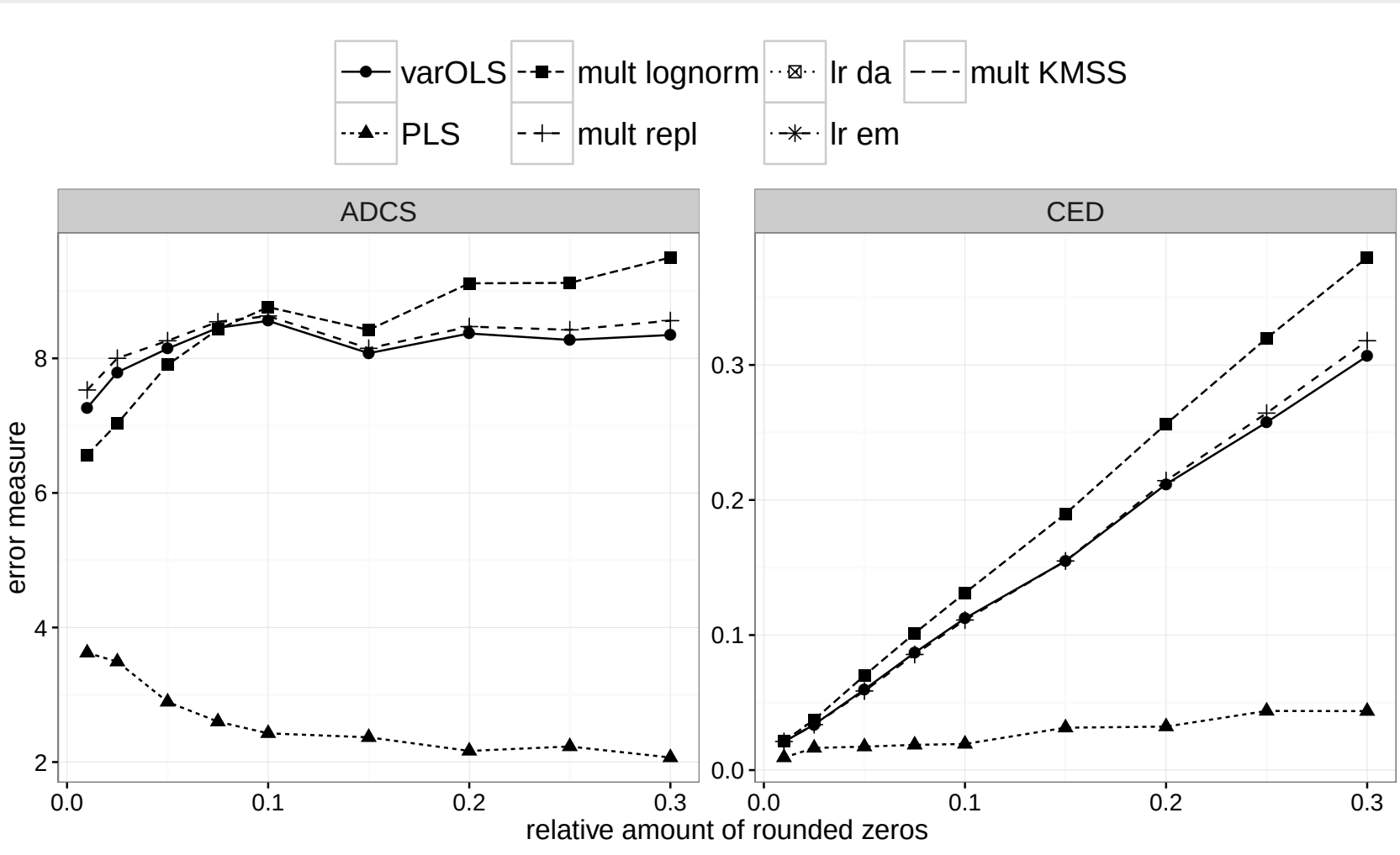
- Normální rozdělení na simplexu (výběrovém prostoru kompozic).

1. **Nízko-dimenzionální scénář**: datová matice $\mathbf{X}$ s $n = 50$ pozorováními a $D = 16$ proměnnými. Podíl hodnot pod DL (nul) je v rozpětí od 0 do 0.3, ty jsou v každé druhé proměnné.
2. **Vysoce-dimenzionální scénář**: datová matice $\mathbf{X}$ s $n = 50$ pozorováními a $D = 128$ proměnnými. Podíl nul stejný jako v nízko-dimenzionálním scénáři.
3. **10% zaokrouhlených nul, rozdílné dimenze**: datová matice $\mathbf{X}$ s $n = 50$ pozorováními a měnícím se počtem prvků kompozice (2, 4, 8, 16, 32, 64, 128, 256).

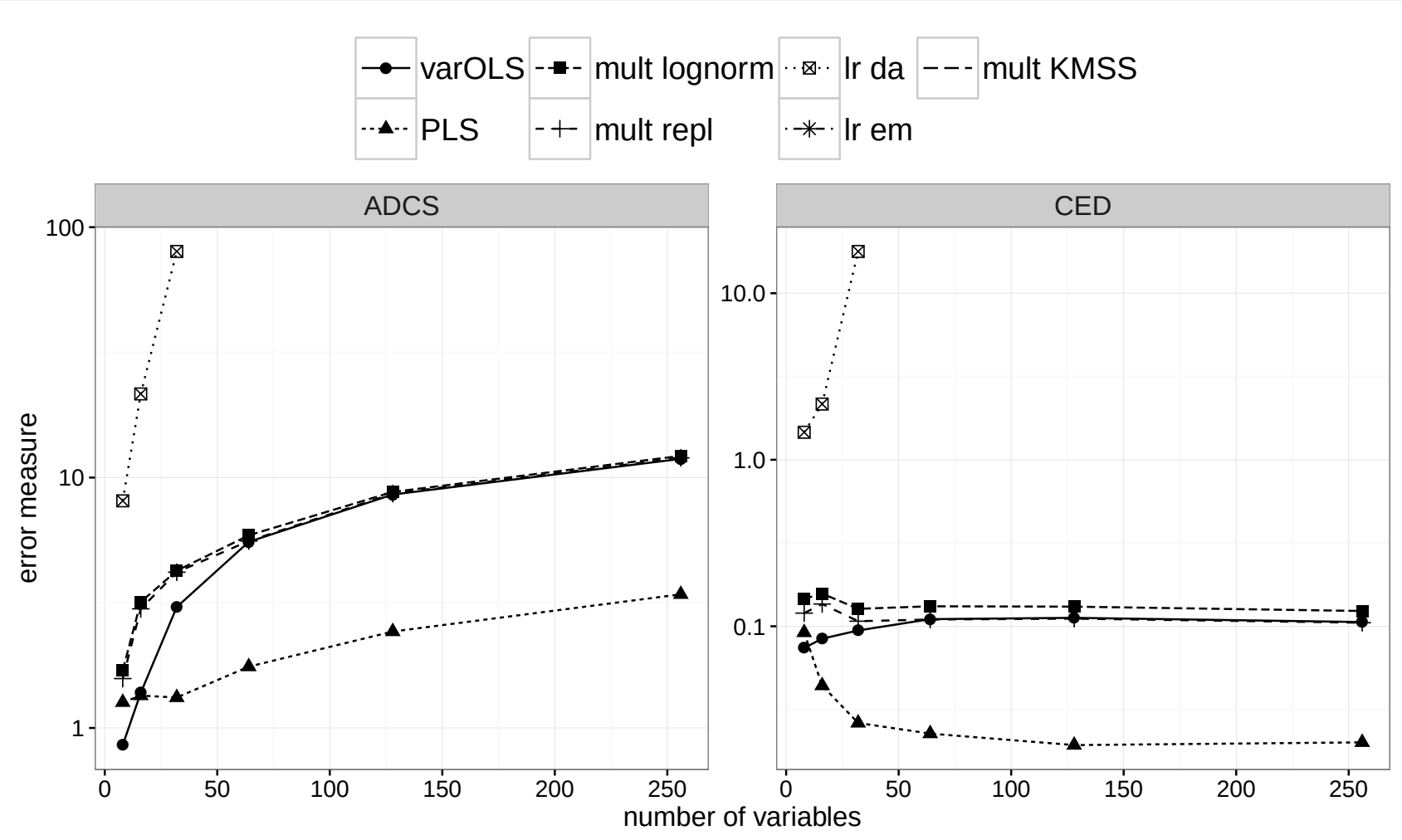
Simulační studie - Nízko-dimenzionální scénář



Simulační studie - Vysoce-dimenzionální scénář



Simulační studie - Rozdílné dimenze



Závěr

- Častý výskyt zaokrouhlených nul v metabolomických datech $\rightarrow$ potřeba jejich imputace.
- Současně používané metody nahrazení (např. použití 2/3 detekčního limitu) nefungují korektně.
- Výhodné použití metody, která kombinuje přístup logratio metodiky a metody dílčích nejmenších čtverců - je zachována mnohorozměrná povaha kompozičních dat.

Literatura


M. Templ, K. Hron, P. Filzmoser, **A. Gardlo**. Imputation of rounded zeros for high-dimensional compositional data. *Chemometrics and Intelligent Laboratory Systems*, 155:183-190, 2016.


J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, London, 1986.


B. Walczak, D.L. Massart. Dealing with missing data. Part I. *Chemometrics and Intelligent Laboratory Systems*, 58:15-27, 2001.


J.A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, J. Palarea-Albaladejo. Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics & Data Analysis*, 56(9):2688-2704, 2012.


K. Hron, M. Templ, P. Filzmoser. Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, 54(12):3095-3107, 2010.


L. Najdekr, **A. Gardlo**, L. Mádrová, D. Friedecký, H. Janečková, E.S. Correa, R. Goodacre, and T. Adam. Oxidized phosphatidylcholines suggest oxidative stress in patients with medium-chain acyl-coa dehydrogenase defficiency. *Talanta*, 139:62-66, 2015.